基于一种改进 Inception 的脱机手写汉字识别 *

陈 站，邱卫根，张立臣

(广东工业大学 计算机学院，广州市 510000)

摘 要：由于字形的复杂多变，脱机手写汉字的识别一直是模式识别的难题，深度卷积神经网络的发展为其提供了一种直接有效的解决方案。研究基于 inceptions 结构神经网络的脱机手写汉字识别,提出了一种 inception 结构的改进方法，它具有结构更加简单、网络深度扩展更加容易、需要的训练参数量更少。该方法在数据集 CISIA-HWDB1.1 上进行了实验验证，采用随机梯度下降优化算法，模型达到了 96.95%的平均准确率。实验结果表明，使用改进的 inception 结构在图像分类上具有更好的鲁棒性，更容易扩展到其他应用领域。

关键词：脱机手写汉字；卷积神经网络；inception

中图分类号：TP391.43 doi: 10.19734/j.issn.1001-3695.2018.09.0784

Offline handwritten Chinese character recognition based on improved inception

Chen Zhan, Qiu Weigen, Zhang Lichen

(*School of Computers Guangdong University of Technology*, *Guangzhou* 510006, *China*)

Abstract: Due to the complexity and variety of glyphs, offline handwritten Chinese character recognition has always been a difficult problem of pattern recognition. The development of deep convolutional neural networks provides a direct and effective solution to this problem. This paper studied offline handwritten Chinese character recognition based on Inceptions neural network.It proposed an improved Inception structure, which took the advantages of simpler structure, easier network depth expansion and less training parameters. The method used the proposed structure to verifiy on dataset CISIA-HWDB1.1. The model achieved an average accuracy of 96.95%, by using stochastic gradient descent optimization algorithm. Experimental result shows that the improved Inception structure has better generalization performance and robustness in image classification, and can be easily extended to other applications.

Key words: offline handwritten Chinese characters; convolutional neural network; inception

0 引言

自 20 世纪 80 年代以来，手写汉字识别(handwritten chinese character recognition, HCCR) 一直是模式识别的一个重要研究领域，也是该研究的难点之一[1]。手写汉字识别的主要困难在于汉字类别数量大、字体结构复杂、字形变化多、书写风格多样，特别是大量相似汉字的存在，使得它们之间的差别极其细微，例如，“已-己”“口-囗”“汩-汨-泪”等，这些高度相似的字符给计算机自动识别带来极大挑战[2]。

经过多年来研究人员的不懈努力，HCCR 取得了极大进展。文献[3]中使用鉴别特征提取方法（discriminative feature learning，DFE）和鉴别学习二次判决函数（discriminative learning quadratic discriminant function，DLQDF）分类器，在脱机手写体汉字数据集 CASIA-HWDB 的几个不同子集上，取得的最好识别率分别是 94.20%（DB1.0）、92.08%（DB1.1）和 92.72% (ICDAR 2013 Competition DB)。

近年来，深度学习逐渐获得了学术界及工业界的广泛重视，在计算机视觉及图像识别领域得到了极其成功的应用，也给手写汉字识别难题带来了新的活力和一些极其有效的解决方法。典型的深度学习结构包括：深度置信网络(DBN)、S 层叠自动编码机(SAE)、卷积神经网络(CNN)、回归神经网络(RNN) 等。近几年，深度卷积神经网络的研究在图像分类上取得了一系列的突破性的进展[4~9]，成为了解决脱机手写汉字识别问题重要工具。文献[10,11]研究基于 CNN 的 HCCR 方法，在 CASIA-HWDB1.0 和 CASIA-HWDB1.1 上都取得不错的结果。CNN 网络结构复杂，全连接层的优化需要庞大的训练数据和计算量。更高的准确率意味着更大的 CNN 网络的深度。同时为了抵消其中必然出现的梯度消失和梯度爆炸的负面影响，网络需要有更复杂的结构。

文献[8,12~14]提出了一种 Inception 结构，并应用于 HCCR。Inception 结构有更小的参数量和更好的鲁棒性。但是，Inception 仍然结构复杂，难以叠加很深的网络深度[9]。本文提出了一种基于改进 Inception 结构的 CNN 网络，为了叙述方便，本文暂称之为 Joint-Net。Joint-Net 不仅具备了 Inception 泛化性能好、参数量小的优点。它在从内部加深网络，提升网络性能的同时，不会产生梯度消失和梯度爆炸的现象。本文的实验表明，它不仅具有比较高的平均准确率，而且容易扩展到其他应用领域。

1 相关工作

Inception 首次提出于 GoogleNet 中[8]，2014 年 Imagenet 竞赛上，22 层的 GoogleNet 取得了冠军，在 ImageNet 数据集上 Top5 错误率达到 6.67%。文献[13]提出了 Inception 的第二个版本，加入了 BN（batch normalization，BN）层，使得模型在 ImageNet 数据集上错误率降低为 4.9%。文献[12]提出了 Inception 的第三个版本，Inception v3 中，$n\times n$ 的卷积核

收稿日期：2018-09-19；**修回日期：**2018-11-22 **基金项目：**国家自然科学基金资助项目（61873068）

作者简介：陈站（1992-），男，硕士，主要研究方向为深度学习、强化学习（1689401214@qq.com）；邱卫根（1968-），男，教授，博士，主要研究方向为人工智能、粗糙集理论及应用、计算机图形图像学；张立臣（1962-），男，教授，博士，主要研究方向为大数据、信息物理融合系统研究.

被拆分成 n×1 和 1×n 两种卷积核，降低网络参数量的同时，提升了网络的识别效果，在 ImageNet 数据集上的 Top5 错误率降低为 3.5%。文献[14]结合了 ResNet[9]网络结构并丰富了 Inception 结构，在 ImageNet 数据集上实现了 3.08%的 Top5 错误率。Inception 的提出与改进大大提升了深度 CNN 网络的识别性能。

Inception 结构的成功，得益于大量使用 1×1 的卷积核和多层次的特征传输。典型的 Inception 结构如图 1 所示。图 1 Inception 结构中，1×1 的卷积核在仅仅增加很小计算量和参数量的情况下，能够增加网络的深度，并改变特征数量，起到升维或者降维的作用。另外，Inception 结构包含了深度为 5，3，2，1 卷积层堆叠的子网络，深度为 5 的子网络大幅增加了网络的深度，而深度为 1 的子网络让特征能更快到达下一个 Inception 结构，缓解了网络深度增大引起的梯度消失和梯度爆炸现象。不同深度的子网络提供了不同层次的特征，这提升了网络对尺度的泛化性能。

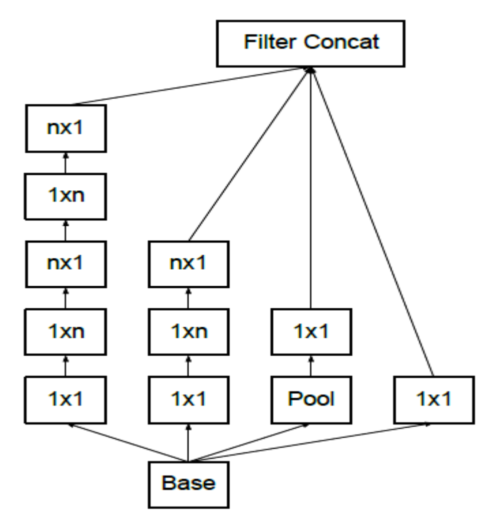


图 1 复杂的 Inception 结构

Fig. 1 Complex Inception structure

Inception 结构复杂不利于堆叠很深的网络，而在梯度消失和梯度爆炸现象未发生时，更深的网络往往有更好的表现[9,15~17]。本文对 Inception 结构进行简化改进，同时保留了 Inception 的优点，使得网络深度的拓展更加容易，同时能提升网络表现。基于改进的 Inception 结构，本文提出 Joint-Net 网络。Joint-Net 大量使用改进的 Inception 结构堆叠网络深度，并去除了最后的全连接层，提升了网络的识别性能的同时，避免了网络的参数量随类别数大幅増加的情况。

2 Joint-Net 网络结构

2.1 改进的 Inception 结构

如图 2 所示，（a）中 Unit 结构由 1×3 卷积核和 3×1 卷积核的卷积层组成，两个卷积的输出按通道拼接一起作为 Unit 的输出。Unit 结构用于取代 3×3 卷积核的卷积层，以便于保留 Inception 结构较小参数量的优点。图 2（b）就是改进后的 Inception 结构。设 Unit 的个数为 N，一方面，（b）结构中包含了深度为 N, N-1, … ,1 的子网络，保留了 Inception 对尺度的适应性。另一方面，对于每一个 Unit，都有来自 Base 的输入和到下一层的直接输出，这种结构能有效避免梯度消失和梯度爆炸现象，因此，在理想情况下，（b）结构的深度是可以非常深的。最后，（b）所有 Unit 的输出按通道拼接后作为 1×1 卷积核的输入，和 Inception 结构一样利用了 1×1 卷积核的优点。

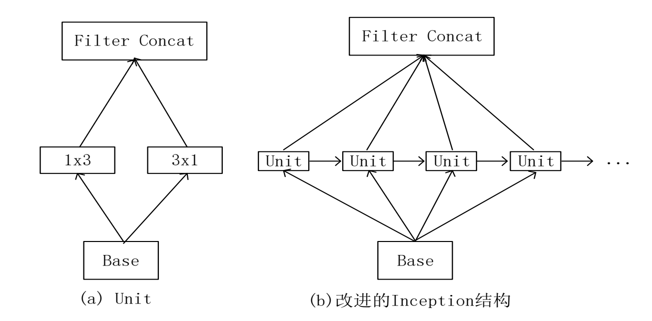


图 2 改进的 Inception 结构

Fig. 2 Improved Inception structure

图 2（b）中，除了第一个 Unit 的输入是来自 Base 之外，每个 Unit 结构的输入都是 Base 和上一个 Unit 结构按通道拼接的结果。除了最后一个 Unit 的输出仅仅直接传向 1×1 卷积层之外，每一个 Unit 结构的输出都复制一份传给下一个 Unit 结构。改进的 Inception 结构保留了 Inception 的优点，增强了网络对尺度的适应性，并使叠加网络深度变得十分方便。

2.2 全卷积的分类模块

深度卷积神经网络的输出层采用独热码（One hot code）对类别进行编码。设分类类别数是 N，类别标签是 n，则对应的编码为

$$A=(0,0,\cdots,0,1,0,\cdots,0)$$

其中：
$$A_i=\begin{cases}1, & i=n\\0, & i\neq n\end{cases}$$

由于输出层采用独热码的编码方式，输出层的人工神经元个数和类别数相同，这导致每层全连接层参数量的空间复杂度是 $O(n^2)$。例如，考虑一级常用汉字 3755 个类别，全连接层输入神经元和输出神经元数都是 3755 个，每个参数占用 4B 空间，则输出层全连接层占用 107MB 以上的空间。而实际应用中，网络可能需要多层全连接层提升识别效果，每层全连接层的参数量也会更大。

本文使用卷积层代替全连接层进行分类。以（特征图个数，特征图高，特征图宽）表示特征图的形状，分类类别数为 n，设最后一层卷积层特征图的形状是 (C,H,W)，则全连接层表示的分类层参数个数为

$$P_{fc}=2nCHW$$

而采用卷积层作为分类层，要求卷积层的输出特征图形状为 (c,h,w)，其中，调整 c 以使 $n=chw$。则以卷积层作为分类层的参数个数为

$$P_{conv}=2n$$

所以，一层全连接层构成分类层的参数量是卷积层构成的分类层的 CHW 倍。

由单层卷积层取代全连接构成输出层会导致网络的识别性能有所下降，而 Joint-Net 对卷积层进行多层卷积层组合调整，使得采用卷积层作为输出层时，分类性能能够达到与使用全连接层同样的效果。

2.3 Joint-Net 网络搭建

改进的 Inception 结构在搭建网络的时候十分方便。只需要多个改进的 Inception 结构直接叠加即可完成网络的主体部分。如图 3 所示，1×1 卷积核的卷积层在网络结构上起衔接作用，本文称之为关节（joint）。在 Inception 内部做池化处理是不方便的，所以在 joint 中加入了可选择的池化层，当需要进行池化时，在 joint 中进行池化。

为使网络模型能有更好的表现，本文将 BN 层和 Relu 层加入了 unit 和 joint，并将 maxpool 层加入到 joint 模块。同

时，由于全部使用1×3 和 3×1 的卷积核组合代替 3×3 卷积核的效果并不好，在加入 3×3 卷积核后，效果优异，所以卷积核增加了 3×3 的选择，方便灵活调整网络，提高网络性能。详细的 unit 和 joint 设计如图 4 所示。

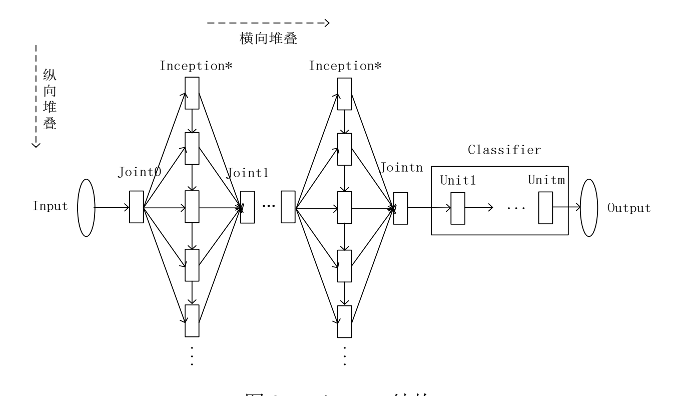


图 3　Joint-Net 结构

Fig. 3　Joint-Net structure

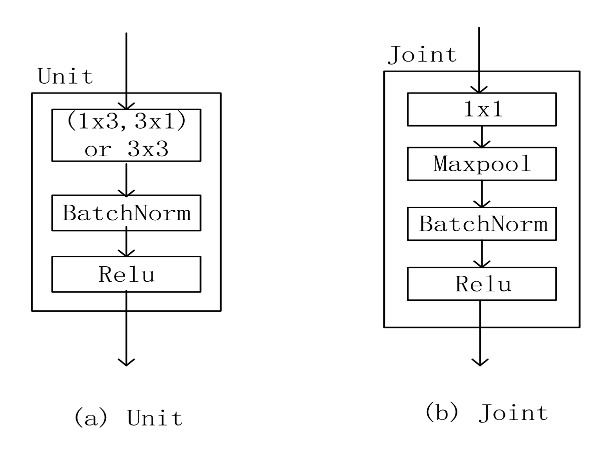


图 4　Unit 与 Joint 的详细结构

Fig. 4　Detailed structure of Unit and Joint

3　实验

为验证 Joint-Net 模型在脱机手写汉字识别上的有效性，本文实验选取了较为大型的 CASIA-HWDB1.1 数据集。脱机手写汉字集 CASIA HWDB1.1 包括了 3755 个 GB2312-80 一级常用汉字。其中，训练集由 240 人手写，测试集由 60 人手写，共计 1121749 个样本，属于大规模模式识别样本集。本文将数据集中的所有图片缩小为 32×32 进行识别实验。

本文所有实验均采用 Pytorch 0.4 在 Windows10 64 位系统上编写及运行代码，实验硬件环境均为 CPU INTER i7 6700K 4.0GHZ，RAM DDR4 8G，GPU GTX1080 8G。

为了验证 Joint-Net 的性能，本文在 CASIA HWDB1.1 上做了大量重复实验。实验采用相同的网络结构，共叠加了 7 个改进的 Inception 结构，8 个关节，和 1 个卷积层。网络结构如表 1 所示，其中 Units 表示相应的 Inception*中的 Unit 部分，Conv*层表示带 Dropout 层的卷积层，用作分类。

本文采用的训练策略为：

a）训练轮数为 60；

b）批大小为 128；

c）学习率调整策略：初始学习率 lr = 0.1，20 轮衰减为 0.02，40 轮衰减为 0.004，50 轮衰减为 0.0008；

d）权重衰减 weight decay = 0.0005；

e）梯度下降：nesterov 加速的 sgd 算法，momentum 为 0.9。

在训练时对样本进行了数据增强，包括边界填充 4 个 0 像素，随机剪切为 32×32 大小，随机水平翻转，归一化。

表 1　网络结构

Table 1　Network structure

	Unit 数量	卷积核大小	输入通道/Unit	输出通道/Unit	池化层
Joint0	1	1x1	1 or 3	32	-
Units1	6	3x3	32	32	-
Joint1	1	1x1	32	64	MaxPool
Unis2	6	3x3	64	64	-
Joint2	1	1x1	64	96	-
Units3	6	3x3	96	96	-
Joint3	1	1x1	96	128	MaxPool
Units4	6	3x3	128	128	-
Joint4	1	1x1	128	160	-
Units5	6	3x3	160	160	-
Joint5	1	1x1	160	192	MaxPool
Units6	6	3x3	192	192	-
Joint6	1	1x1	192	224	-
Units7	6	3x3	224	224	-
Joint7	1	1x1	224	512	MaxPool
Conv*	1	1x1	512	N/(2x2)	-

图 5 展示了 Joint-Net 在 CASIA HWDB1.1 训练集和测试集上的训练情况。在图 5 中，Joint-Net 在 CASIA HWDB1.1 的训练集上进行训练，在 CASIA HWDB1.1 测试集上进行测试。可以看出，在每次学习率衰减时，准确率都有明显的提高，最终网络收敛。实验结果与其他网络模型结果的对比如表 2 所示，其中，带“*”的表示文献没有单独在 CASIA HWDB1.1 数据集上进行训练，本文复现实验取得的结果。

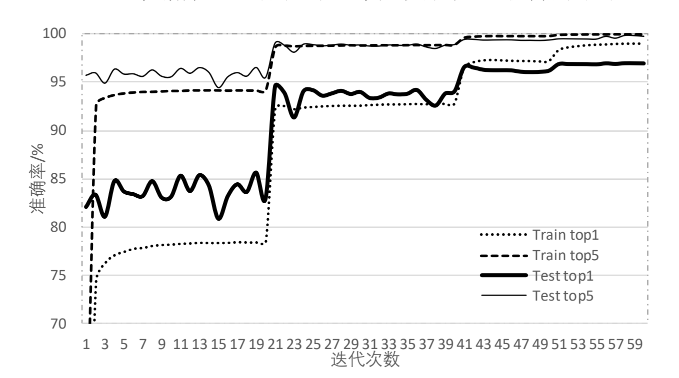


图 5　Joint-Net 网络模型准确率变化过程

Fig. 5　Joint-Net network model accuracy rate change process

由于去除了全连接层，尽管 Joint-Net 网络有更深的卷积层数，样本集的类别多达 3755 个，模型的参数量仍然很小，而且参数量不会随类别的增加迅速增大。同时模型有更好的鲁棒性和泛化能力，在 CASIA HWDB1.1 数据集上比文献[10,11]的结果有显著的提升。

为验证 Joint-Net 的实用性，本文将 Joint-Net 网络应用于汉字字幕提取系统，达到了很高的识别率，表明 Joint-Net 适宜在实际系统上取得应用。汉字字幕提取模型图 6 所示。

图 6 中，先由预训练好的 Joint-Net 模型对带有汉字字幕的图片进行处理，得到包含 3756 个类别得分（包含 1 个背景类别和 3755 个常用汉字类别）。由于不关注字幕的位置，仅提取字幕信息，所以没有进行坐标回归。图 6 中，为说明 Infer 的过程，记 Class scores 为张量 X，Class map 为张量 Y，maxpool 是尺寸为 3×3，步幅为 1 的最大池化操作，argmax 为求取像素点最大值所在通道的操作，则 Infer 过程为

$$X^* = maxpool(X)$$

$$Y = \mathrm{argmax}(\mathrm{relu}(X - X^*))$$

由于图 6 中 Infer 操作对类别得分取的是局部的最大值，相当于进行了多次投票得出的类别，准确率比单一汉字识别要高。但是由于对每个像素为中心的 5×5 的区域都进行分类，速度有所降低。采用图 6 的模型，在 200 张包含汉字的实际视频截取视频帧中的测试中，召回率（recall）达到了 98.9%，准确率（accuracy）为 98.4%，速度为 14 张/s。

表 2 Joint-Net 模型在 CASIA HWDB1.1 上的准确率与其他网络模型的对比

Table 1 Compared to other network models

network	params.	CASIA HWDB1.1(accuracy %)
DirectMap+ConvNet+Adaptation[10]	23.5M	96.55
*HCCR-CNN12Layer+GSLRE 4X[11]	32.7M	96.73
Joint-Net	**8.1M**	**96.95**

图 6 使用 Joint-Net 的汉字字幕提取系统模型

Fig. 6 Using the Chinese character subtitle extraction system model of Joint-Net

4 结束语

实验结果显示，Joint-Net 在包含 3755 个类别的脱机手写汉字集 CASIA HWDB1.1 上达到了公开报告的单模型的最佳成绩。这表明 Joint-Net 是一种学习能力强大且鲁棒性和泛化能力优异的卷积神经网络结构。其独特的单元层和关节，能有效的增加网络的深度，并提升网络的鲁棒性和泛化能力，使网络更容易的达到更好的结果。实验中，在实际汉字字幕提取系统中的成功应用表明 Joint-Net 模型具有一定的实用价值。由于 Joint-Net 结构简单的特性，能够很容易的将网络模型压缩算法[1][8]应用到模型中，这将是对 Joint-Net 进一步研究的方向。

参考文献：

[1] 赵继印，郑蕊蕊，吴宝春，等. 脱机手写体汉字识别综述 [J]. 电子学报, 2010, 38 (2): 405-415. (Zhao Jiyin, Zheng Ruirui, Wu Baochun, *et al.* A review of offline handwritten Chinese character recognition [J]. Acta Electronica Sinica, 2010, 38 (2): 405-415.)

[2] 金连文，钟卓耀，杨钊，等. 深度学习在手写汉字识别中的应用综述 [J]. 自动化学报，2016, 42 (8): 1125-1141. (Jin Lianwen, Zhong Zhuoyao, Yang Zhao, *et al.* Applications of deep Learning for handwritten Chinese character recognition: a review [J]. Acta Automatica Sinica, 2016, 42 (8): 1125-1141.)

[3] Liu Chenglin, Yin Fei, Wang Dahan, *et al.* Online and offline handwritten Chinese character recognition: benchmarking on new databases [J]. Pattern Recognition, 2013, 46 (1): 155-162.

[4] LeCun Y, Boser B, Denker J S, *et al.* Backpropagation applied to handwritten zip code recognition [J]. Neural computation, 1989, 1 (4): 541-551.

[5] LeCun Y, Bottou L, Bengio Y, *et al.* Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL].(2015-04-10). https://arxiv.org/abs/1409.1556.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classifica-tion with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. 2012: 1097-1105.

[8] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going deeper with convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society,2015: 1-9.

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society,2016: 770-778.

[10] Zhang Xuyao, Bengio Y, Liu Chenglin. Online and offline handwritten Chinese character recognition: a comprehend-sive study and new benchmark [J]. Pattern Recognition, 2017, 61: 348-360.

[11] Xiao Xuefeng, Jin Lianwen, Yang Yafeng, *et al.* Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition [J]. Pattern Recognition, 2017, 72: 72-81.

[12] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the Inception architecture for computer vision [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society,2016: 2818-2826.

[13] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]//Proc of International Conference on Machine Learning. 2015: 448-456.

[14] Szegedy C, Ioffe S, Vanhoucke V, *et al.* Inception-v4, Inception-ResNet and the impact of residual connections on learning [EB/OL]. (2016-08-23). https://arxiv.org/abs/1602.07261.

[15] Srivastava R K, Greff K, Schmidhuber J. Highway networks [J]. arXiv preprint arXiv: 1505. 00387, 2015.

[16] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks [C]//Advances in Neural Information Processing Systems. 2015: 2377-2385.

[17] Huang Gao, Liu Zhuang, Van Der Maaten L, *et al.* Densely connected convolutional networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society, 2017.

[18] Hu Jie, Shen Li, Albanie S, *et al.* Squeeze-and-excitation networks [EB/OL]. (2018-10-25). https://arxiv.org/pdf/1709.01507.pdf.